An Overview on Data Quality Issues at Data Staging ETL

Nitin Anand [1] and Manoj Kumar[2]
[1]Research Scholar , Deptt of Computer Science , AIACT&R, New Delhi
Email: proudtobeanindiannitin@gmail.com
[2]Associate Professor, Deptt of Computer Science , AIACT&R, New Delhi
Email: manojgaur@yahoo.com

***Abstract* -A data warehouse (DW) is a collection of technologies aimed at enabling the decision maker to make better and faster decisions. Data warehouses differ from operational databases in that they are subject oriented, integrated, time variant, non volatile, summarized, larger, not normalized, and perform OLAP. The generic data warehouse architecture consists of three layers (data sources, DSA, and primary data warehouse). During the ETL process, data is extracted from an OLTP databases, transformed to match the data warehouse schema, and loaded into the data warehouse database**

***Index Terms* - Data Mart, Data Quality (DQ), Data Staging , Data Warehouse, ETL, OLAP,OLTP.**

I. Introduction

Extraction-Transformation and Loading (ETL) tools are pieces of software responsible for the extraction of data from several sources, their cleansing, customization and insertion into a data warehouse. The quality of the information depends on 3 things: (1) the quality of the data itself, (2) the quality of the application programs and (3) the quality of the database schema ETL and data staging is considered to be more crucial stage of data warehousing process where most of the data cleansing and scrubbing of data is done. There can be myriad of reasons at this stage which can contribute to the data quality problems. To build a DW we must run the ETL tool which has three tasks: (1) data is extracted from different data sources, (2) propagated to the data staging area where it is transformed and cleansed, and then (3) loaded to the data warehouse. ETL tools are a category of specialized tools with the task of dealing with data warehouse homogeneity, cleaning, transforming, and loading problems [1].The preparation of data before their actual loading in the warehouse for further querying is necessary due to quality problems, incompatible schemata, and unnecessary parts of source data not relevant for the purposes of the warehouse.

The category of tools that are responsible for this task is generally called Extraction- Transformation- Loading (ETL) tools. The functionality of these tools can be coarsely summarized in the following prominent tasks, which include:
1. The identification of relevant information at the source side.
2. The extraction of this information,
3. The customization and integration of the information and integration of the information coming from multiple sources [2].
4. The cleaning of the resulting data set on the basis of database and business rules, and
5. The propagation of the data to the data warehouse and/ or data marts.

II. Literature Review

1. Amit Rudra and Emilie Yeo (1999)"Key Issues in Achieving Data Quality and Consistency in Data Warehousing among Large Organizations in Australia
2. Munoz, Lilia, Mazon, Jose-Norberto, Trujillo, Juan, 2010. Systematic review and comparison of modeling ETL processes in data warehouse.
3. Jaideep Srivastava Warehouse Creation- A Potential Roadblock to Data Warehousing.

III. Phases of Etl

An ETL system consists of three consecutive functional steps: extraction, transformation, and loading:

A. Extraction

The ETL Extraction step is responsible for extracting data from the source systems. Each data source has its distinct set of characteristics that need to be managed in order to effectively extract data for the ETL process. The process needs to effectively integrate systems that have different platforms, such as different database management systems, different operating systems, and different communications protocols.

B. Transformation

The second step in any ETL scenario is data transformation. The transformation step tends to make some cleaning and con-forming on the incoming data to gain accurate data which is correct, complete, consistent, and unambiguous. This process includes data cleaning, transformation, and integration. It de-fines the granularity of fact tables, the dimension tables, DW schema (stare or snowflake), derived facts, slowly changing fact tables and dimension tables . All transformation rules and the resulting schemas are described in the metadata repository.

C. Loading

Loading data to the target multidimensional structure is the final ETL step. In this step, extracted and transformed data is written into the dimensional structures actually

DOI: 03.LSCS.2013.3.47

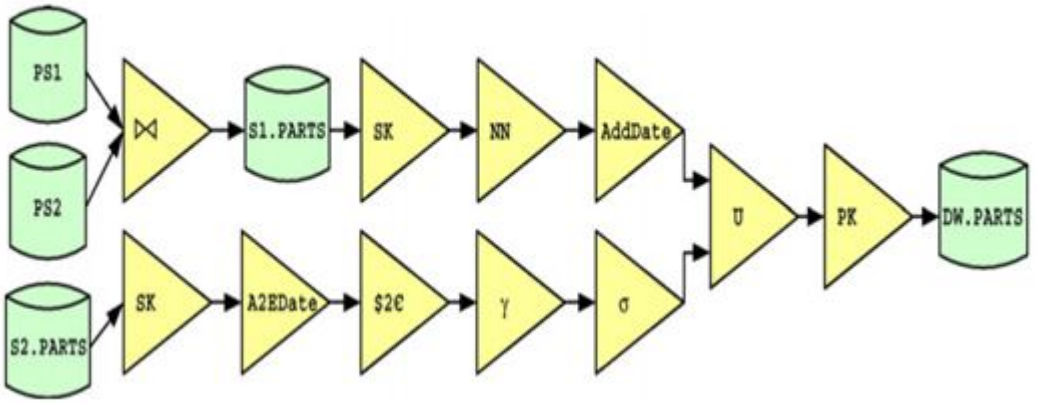

Fig 1 ETL workflow as a directed graph [19]

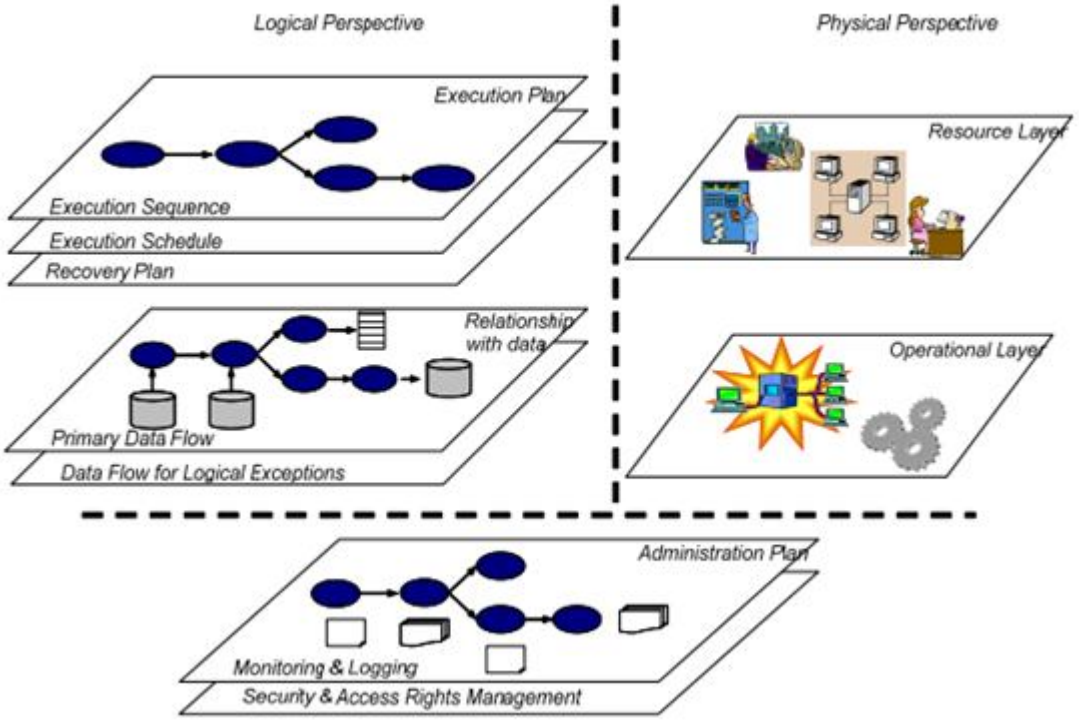

Fig. 2 Different perspectives for an ETL workflow [16]

accessed by the end users and applications.

IV. A Rationale For The Taxonomy

An ETL workflow can be seen as a directed graph as shown in Figure 1. The nodes of this graph are activities and recordsets [17]. The edges of the graph are relationships that combine activit ies and recordsets.

The edges of the graph are provider relationships that combine activities and recordsets [3]. Following the common practice, we envisage ETL activities to be combined in a workflow.

Therefore, we do not assume that the output of a certain activity will be necessarily directed towards a recordset, but rather, that the recipient of this data can be either another activity or a recordset.[16]. In Figure 2 [10]

We follow a multi-perspective approach that enables to separate these parameters and study them in a principled approach [16]. We are mainly interested in the design and administration parts of the lifecycle of the overall ETL process, and we depict them at the upper and lower part of Fig. 2, respectively. At the top of Fig. 2, we are mainly concerned with the static design artifacts for a workflow systems. These loading steps includes both loading dimension tables and loading fact tables [4]. We will follow a traditional approach and group the design artifacts into physical, with each category comprising its own perspective. We depict the logical perspective on the left-hand side of Fig. 2, and the physical perspective on the right-hand side. At the logical perspective, we classify the design artifacts that give an abstract description of the workflow environment. First, the designer is responsible for defining an execution plan for the scenario. The definition of an execution plan can be seen from various perspectives. The execution sequence involves the specification of which activity runs first, second, and so on, which activities run in parallel, or when a semaphore is de

DOI: 03.LSCS.2013.3.47

fined so that several activities are synchronized at a rendezvous point. ETL activities normally run in batch, so the designer needs to specify an execution schedule, i.e., the time points or events that trigger the execution of the scenario as a whole. Finally, due to system crashes, it is imperative that there exists a recovery plan, specifying the sequence of steps to be taken in the case of failure for a certain activity (e.g., retry to execute the activity, or undo any intermediate results produced so far). On the right-hand side of Fig. 2, we can also see the physical perspective, involving the registration of the actual entities that exist in the real world. We will reuse the terminology of [5] for the physical perspective. The resource layer comprises the definition of roles (human or software) that are responsible for executing the activities of the workflow. The operational layer, at the same time, comprises the software modules that implement the design.

V. Causes Of Data Quality Issues At Data Staging Etl Phase

One consideration is whether data cleansing is most appropriate at the source system, during the ETL process, at the staging database, or within the data warehouse [6] [7]. A data cleaning process is executed in the data staging area in order to improve the accuracy of the data warehouse. The data staging area is the place where all 'grooming' is done on data after it is called from the source systems. Staging and ETL phase is considered to be most crucial stage of data warehousing where maximum responsibility of data quality efforts resides. Table 1 depicts some reasons for this [18]

VI. Algorithm For Simulation

The activity diagram (shown in figure 3) depicts the method of secure data extraction. The main procedure for secure data extraction process [12,13] is as follows.
// Identifying the sources and creating the source list.
This is done by the methods of Source Identifier class
1. Identify the list of clients attached to the server
2. Find the type of the databases by pinging to that client
3. Set the properties for the source
4. If it is a new source add to the data source list // Establishing the connection and extracting data.
This is done by methods of Wrapper class
5. Check the type of the data source
6. Using appropriate drivers establish the connection
7. Map the data source and data staging area schemas
8. Extract the data // Loading of extracted data into data staging area. Integrator
class does this.
9. Establish connection with data staging area
10. Install the data into data staging area // Modification / updation of Data Staging Area (DSA).
Integrator updates DSA with the help of Monitor.
11. Identify the changes in the data sources and Inform to the Integrator
12. Update DSA

Table I: Data Quality issues at ETL

Sl no	CAUSES OF DATA QUALITY ISSUES AT DATA STAGING AND ETL PHASE.
1	Business rules lack currency problems [8]
2	Lack of capturing only changes in source files [9]
3	Disabling data integrity constraints in data staging tables cause wrong data and relationships to be extracted and hence cause data quality problems [10].
4	The inability to restart the ETL process from checkpoints without losing data [11]
5	Lack of Providing internal profiling or integration to third-party data profiling and cleansing tools.[11]
6	Lack of automatically generating rules for ETL tools to build mappings that detect and fix data defects[11]
7	Inability of integrating cleansing tasks into visual workflows and diagrams[11]
8	Inability of enabling profiling, cleansing and ETL tools to exchange data and meta data[11]
9	Lack of proper functioning of the extraction logic for each source system (historical and incremental loads) cause data quality problems.
10	Lack of generation of data flow and data lineage documentation by the ETL process causes data quality problems
11	Lack of error reporting, validation, and metadata updates in ETL process cause data quality problems.
12	Inappropriate handling of rerun strategies during ETL causes data quality problems.
13	Lack of considering business rules by the transformation logic cause data quality problems
14	Wrong impact analysis of change requests on ETL cause data quality problems.
15	Type of staging area, relational or non relational affects the data quality
16	The inability to schedule extracts by time, interval, or event cause data quality problems
17	Hand coded ETL tools used for data warehousing lack in generating single logical meta data store, which leads to poor data quality.

VII. Conclusion And Futurework

In this paper we have attempted to collect all possible causes of data quality problems that may exist at all the phases of data warehouse. In a recent study [14], the authors report that due to the diversity and heterogeneity of data sources, ETL is unlikely to become an open commodity market. This paper describes the simulation model of Secure Data Extraction in ETL processes. This architecture gives us flexibility of adding various types of information sources, which ultimately helps in storing the data into the Data Staging Area. Since quality plays an important role in developing software products, I have presented functional requirements along with non-functional requirement i.e., security requirements.. This approach is better as compared to existing systems. In [15], the authors report on their data warehouse population system. The architecture of the system is discussed in the paper, with particular interest (a) in a "shared data area which is an in-memory area for data transformations, with a specialized area for rapid access to lookup tables and (b) the pipelining of the ETL processes. Our classification of causes

DOI: 03.LSCS.2013.3.47

Figure 3: Activity diagram for data extraction

will really help the data warehouse practitioners, implementers and researchers for taking care of these issues before moving ahead with each phase of data warehousing. Each item of the classification shown in table, will be converted into a item of the research instrument such as questionnaire and will be empirically tested by collecting views about these items from the data warehouse practitioners, appropriately.

REFERENCES

[1] Shilakes, C., Tylman, J., 1998. Enterprise Information Portals. Enterprise Software Team. <http://www.sagemaker.com/company/downloads/eip/indepth.pdf>.

[2] J. Adzic and V. Fiore, "Data Warehouse Population Platform," Proc. Fifth Int l Workshop Design and Management of Data Warehouses, 2003.

[3] A. Simitsis, P. Vassiliadis, and T. Sellis, "Optimizing ETL Processes in Data Warehouses," Proc. 21st IEEE Int'l Conf. Data Eng.,pp. 564-575, 2005.

[4] Panos Vassiliadis, Alkis Simitsis, Panos Georgantas, Manolis Terrovitis and Spiros Skiadopoulos"A Generic and customizable framework for the design of ETL scenarios",2002

[5] W.M.P. van der Aalst, A.H.M. ter Hofstede, B. Kiepus-zewski, A.P. Barros. Workflow Patterns, BETA Working Paper Series, WP 47, Eindhoven University of Technology, Eindhoven, 2000, available at the Workflow Patterns website, at tmithttp:/ /www.tm.tue.nl/research/patterns/documentation.htm

[6] Wayne W. E. (2004) "Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data ", The Data warehouse Institute (TDWI) report , available at www.dw-institute.com .

[7] Ralaph Kimball, The Data Warehouse ETL Toolkit, Wiley India (P) Ltd (2004)

[8] Amit Rudra and Emilie Yeo (1999)"Key Issues in Achieving Data Quality and Consistency in Data Warehousing among Large Organizations in Australia", Proceedings of the 32nd Hawaii International Conference on System Sciences – 1999

[9] Arkedy Maydanxhik (2007), " Causes of Data Quality Problems", Data Quality Assessment, Techniques Publications LLC. Available at http://media.techtarget.com/searchDataManagement/downloads/Data_Quality_Assessment_-_Chapter_1.pdf

[10] Won Kim et al (2002)- "A Taxonomy of Dirty Data " Kluwer Academic Publishers 2002

[11] Wayne Eckerson & Colin White (2003) " Evaluating ETL and Data Integratio Vassiliadis A generic and customizable framework for the design of ETL scenarios, 2002

[17] Panos Vassiliadis, A Taxonomy of ETL activities 2009.

[18] Ranjit Singh, Dr Kanwaljeet, A descriptive classification of causes of Data Quality problems in Data Warehouse,2010

[19] A.Simitsis. Mapping Conceptual to Logical Models for ETL Processes . DOLAP 05, ACM, (2002) 67-76

DOI: 03.LSCS.2013.3.47